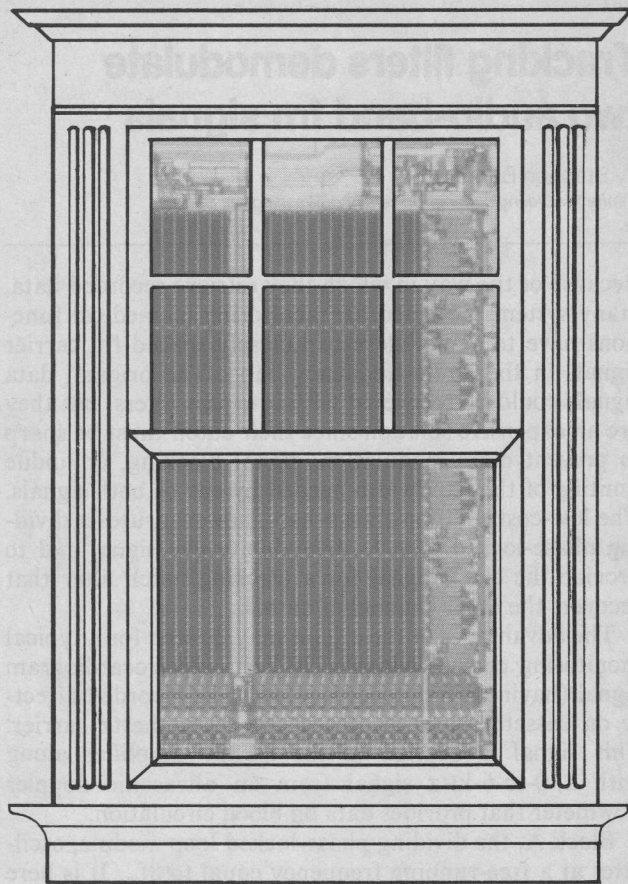


The same quartz window that lets in the ultraviolet light for erasure also permits the user to see the actual silicon chip. Maybe this has something to do with the enormous popularity enjoyed by the E-PROM, or erasable programmable read-only memory. But more likely responsible is its attractive combination of field programmability, high density, and low power consumption when compared to bipolar fusible-link memories. Many an E-PROM has been shipped in end equipment because of these desirable features.

The same company that pioneered the E-PROM in the early 1970s wrote this two-part tutorial on its fabrication and reliability. Part 1 reviews the structure of the E-PROM's floating-gate memory cell and the mechanisms of its programming and erasure. As this section notes, electrons accelerated in the channel of the memory transistor gain enough energy during programming to penetrate an oxide dielectric and to cling to a floating gate, made attractive with an overlapping, capacitively coupled control gate. These captive electrons scatter from the floating gate, however, when hit with a flood of UV radiation.

The E-PROM structure and operation give rise to some special reliability considerations (part 2). For example, the oxide insulator surrounding the floating gate must have the integrity to hold charge yet be thin enough to pass electrons and light. Various failure modes of the E-PROM are introduced and discussed in relation to the lifespan of the chip. Specific reliability data is presented for the 16-K 2716 E-PROM, and screening procedures are outlined that can be used to weed out bad devices and estimate total failure rate for each batch.

-John G. Posa



E-PROM reliability: part 1

An E-PROM's integrity starts with its cell structure

A practical understanding of avalanche injection, floating gates, and UV erasure has led to predictable device behavior

by Murray H. Woods, Intel Corp., Santa Clara, Calif.

□The ultraviolet-light-erasable programmable read-only memory has matured, and its processing become rather routine. This situation was long in arriving. For years, in fact, the right sequence of fabrication steps was elusive because of the E-PROM's many potential failure modes. Owing to the unique structure of the E-PROM, these afflictions can strike during programming, erasing, or reading.

The E-PROM transistor resembles an ordinary MOS transistor, except for the addition of a floating gate, buried in the insulator between the substrate and the ordinary select-gate electrode (Fig. 1). As a result, in an

E-PROM, the select-gate voltage must be capacitively coupled in series with the floating gate rather than directly to the underlying channel. Charge stored on the floating gate alters the threshold voltage of the device as seen by the top or select gate.

Figure 1 also shows why very high densities can be achieved with E-PROM cells: the floating storage gate and the select gate are both directly above the transistor's channel. Even higher densities are obtained by self-aligning these two gates and the source and drain regions. This vertical stacking allows the functions of storage and reading—which normally require separate

devices—to reside within the space a single field-effect transistor occupies.

The cell is programmed by charging the floating gate via the injection of so-called hot electrons from the drain's pinch-off region (Fig. 2a). It is erased through internal photoemission from the floating gate to the top gate and substrate. The ultraviolet light gives electrons on the floating gate enough energy to surmount the energy barrier between the floating gate and the insulator surrounding it (Fig. 2b).

The charge on the memory cell's floating storage gate changes the threshold voltage of the select gate by an amount $\Delta V_t = -\Delta Q_{FG}/C$, where C is the capacitance between the floating gate and the select gate and ΔQ_{FG} is the change in charge on the floating gate. The drain current versus select-gate voltage transfer characteristics for the programmed and erased states are parallel to each other (Fig. 3). The select-gate voltage during reading lies between these two curves and results in a drain current that reflects the cell's state, the nonconducting programmed state (storing a 0), or the conducting erased state (storing a 1). When the E-PROM cell is programmed, the negative charge on the floating gate causes the floating-gate-to-source voltage to be negative. This turns the cell off, even with a positive reading voltage applied to the select gate.

Since the floating gate is not tied to a power supply, its voltage is determined by its charge and by capacitive coupling to the voltages of the select gate, the drain, the channel, and the source. The difference between the floating-gate voltage and the voltages of these other cell areas can be used to determine the electric fields in the various oxide regions of the device.

Although the voltage applied to the chip during programming is high—25 volts—the fields across the gate oxide directly above the channel are relatively small, about 2 megavolts per centimeter at the onset of programming and less than 1.5 MV/cm afterward. During all other operations, even when the floating gate is fully charged, this field is ≤ 0.7 MV/cm. This is a typical operating field for a high-performance MOS device.

Programming

The hot electrons get their energy from the voltage applied to the drain of the E-PROM cell. They are accelerated along the channel into the even higher fields surrounding the drain depletion region. While traversing the channel, the electrons enter a region where the electric field in the substrate is about 10^5 V/cm or greater. At this point the rate of energy gained from the electric field can no longer be described by the temperature of the silicon; hence the term "hot." Once these electrons gain sufficient energy they can surmount the energy barrier of about 3.2 electronvolts between the silicon substrate and the silicon dioxide insulator. Because energy loss due to phonon emission increases at higher lattice temperatures, it is actually easier to obtain hot electrons at lower operating temperatures.

In addition to phonon emission, hot electrons with energies above about 1.8 eV may give up some of this energy in another way: through electron-hole pair creation resulting from impact ionization. This phenomenon

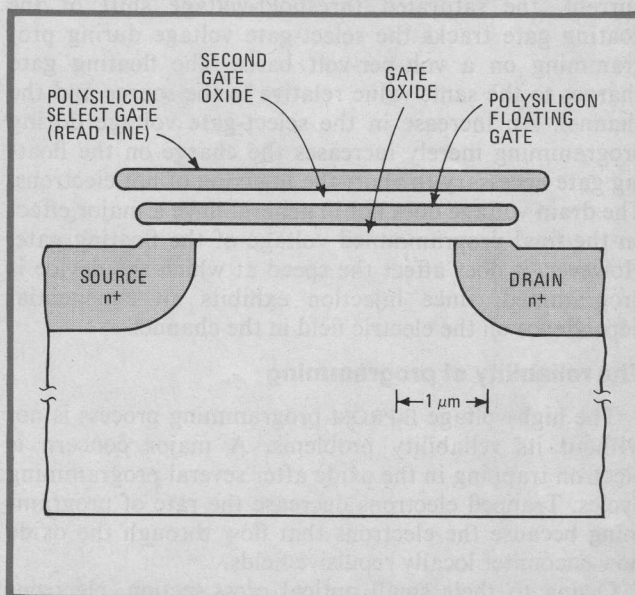
is observed in ordinary MOS transistors as the cause of the onset of substrate current at high drain voltages. However, in the case of the E-PROM, significant current multiplication produces substantial substrate current even before a large enough drain voltage is reached to produce hot-electron injection into the oxide.

With positive drain and channel voltages, electrons injected into the oxide of an n-channel E-PROM return to the substrate unless a high positive select-gate voltage is applied to pull the electrons toward the floating gate. Not only does the floating gate have to be positively biased with respect to the source, it must also be positive with respect to the point along the channel where hot-electron injection occurs.

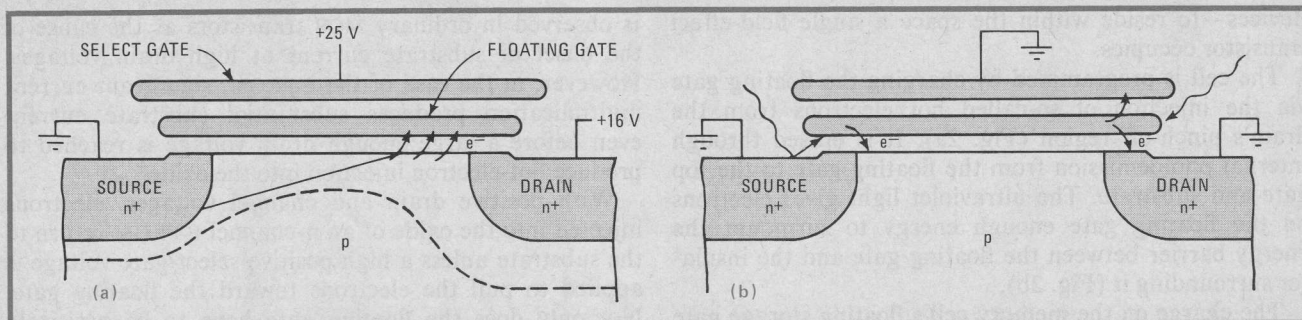
Near the beginning of the injection process, the inversion layer extends almost all the way to the drain, and the field in the oxide is attractive except for a small portion very near the drain (see Fig. 4). Current begins to flow through the oxide at the point where the electrons are their hottest and where the oxide field is most favorable. As the floating gate charges up, the floating-gate-to-source voltage drops and the drain's pinch-off region moves towards the source. The surface field near the drain intensifies and more hot electrons are produced in the substrate.

However, as seen in Fig. 4, in the region where the electrons are their hottest, the oxide field is least favorable for injection and so the injected-electron current begins to subside. Thus, fortunately, the electron injection process is self-limiting. The charging of the floating gate reduces the number of electrons that can be accelerated in the high-field region. As the floating gate becomes fully charged, the oxide current is reduced almost to zero because the oxide field is now repulsive to the electrons injected into the high-field region.

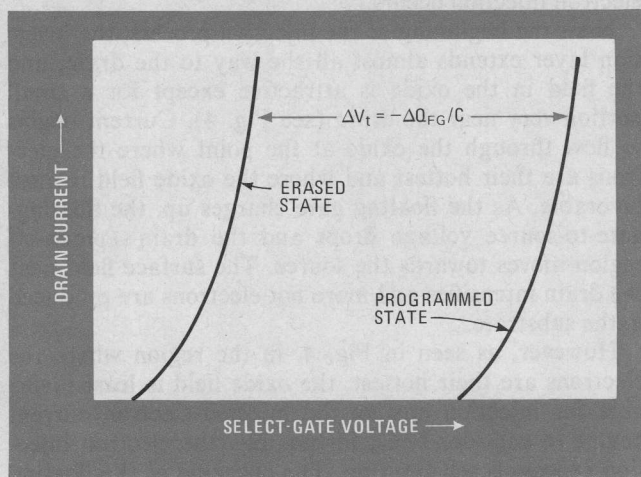
Since the reducing electric field between the floating gate and channel is responsible for shutting off the oxide



1. Vertical. In the erasable programmable read-only memory, a floating storage gate and a select gate are directly above a MOS FET's channel. The floating gate assumes the select-gate potential during writing. The 1-micrometer scale applies to the 2716.



2. To and fro. During programming of the E-PROM cell, electrons are injected onto a floating polysilicon gate (a). During erasure the same electrons receive enough energy from ultraviolet radiation to enter the surrounding silicon dioxide layer (b).



3. Sensible. An E-PROM cell's threshold voltage determines whether it is sensed as a logic 0 (the nonconducting programmed state) or 1 (erased). The change in threshold corresponds to the shift shown in the select-gate voltage to drain current transfer characteristic. This shift depends on the difference in charge on the floating gate in the two states (ΔQ_{FG}), as well as the intergate capacitance (C).

current, the saturated threshold-voltage shift of the floating gate tracks the select-gate voltage during programming on a volt-per-volt basis. The floating gate charges to the same value relative to the source and the channel. An increase in the select-gate voltage during programming merely increases the charge on the floating gate necessary to abort the injection of hot electrons. The drain voltage does not in general have a major effect on the final programmed voltage of the floating gate. However, it does affect the speed at which the device is programmed, since injection exhibits an exponential dependence on the electric field in the channel.

The reliability of programming

The high-voltage E-PROM programming process is not without its reliability problems. A major concern is electron trapping in the oxide after several programming cycles. Trapped electrons decrease the rate of programming because the electrons that flow through the oxide now encounter locally repulsive fields.

Owing to their small optical cross section, electrons can easily remain trapped in the oxide even after UV erasure of the floating gate. Therefore, as a result of the field around them, trapped electrons also locally raise the threshold voltage of the channel. Indeed, the erased

threshold voltage can, through trapping, rise to a point where cell states are sensed incorrectly. This condition is avoided by proper growth of the first gate oxide to reduce the density of electron traps.

Another programming problem may arise when the select-gate voltage is raised high for a device in a row that contains another device that has already been programmed. This produces the highest field in the upper, or second, gate oxide during any time of operation. The average electric field in the second oxide is only about 1 MV/cm, but pointed asperities on the first polysilicon layer can substantially increase the local electric field to the point where partial erasure of the floating gate occurs. This field emission effect can be eliminated by the right combination of process steps, including the temperature used to form the gate structure.

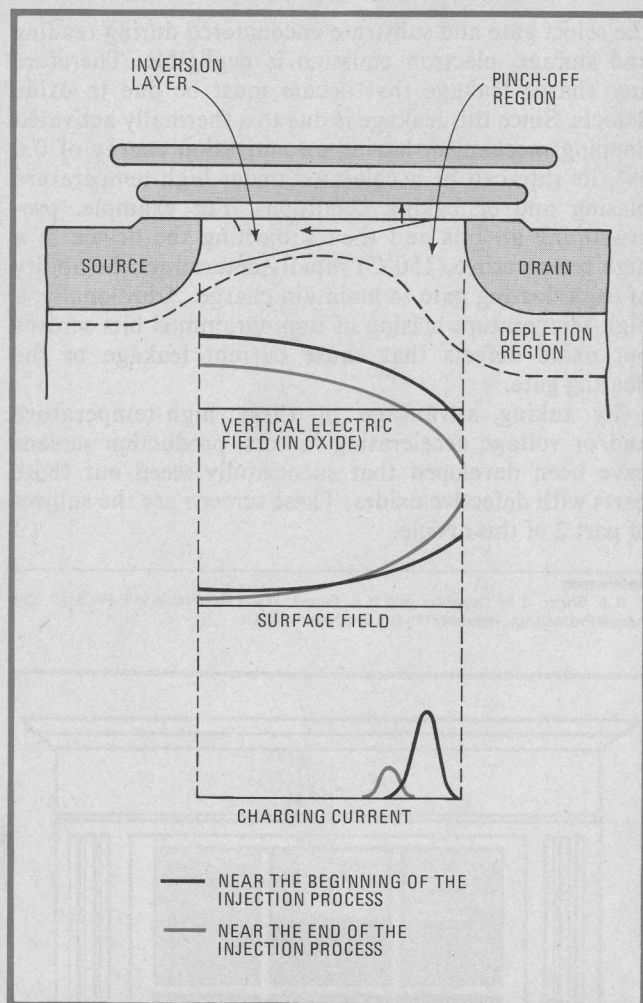
Another case of erasure through field emission may occur when a column is raised high to program a device in another row. The drains of the unselected programmed cells in that column will rise in voltage to create a field of approximately 2.5 MV/cm in the first gate oxide, that between the drain and the floating gate. Fortunately, the bottom of the polysilicon floating gate has a smooth surface so tunneling (Fowler-Nordheim emission) of electrons from the floating gate to the substrate is negligible.

Erase conditions

Complete erasure of an E-PROM is required before the device can be reprogrammed. This is accomplished by exposing the entire array to UV light. Typical sources for erasure are quartz-jacketed mercury arc lamps and mercury vapor lamps, which emit strong radiation with a wavelength of 2,537 angstroms (4.9 eV).

The photons are absorbed by electrons in the conduction and valence bands of the polysilicon floating gate; at this wavelength, most are absorbed within 50 Å of the oxide interface. The excited electrons leave the polysilicon floating gate, enter the oxide, and are swept away to the select gate or substrate by the local field. During erasure, the select gate, source, drain, and substrate are all near ground potential.

With an n-type polysilicon floating gate, electrons can be excited from either the conduction band or the valence band into the oxide. Excitation from the conduction band requires only 3.2 eV, while the barrier height from the valence band is 4.3 eV. Even for heavily doped n-type material there are many more electrons available

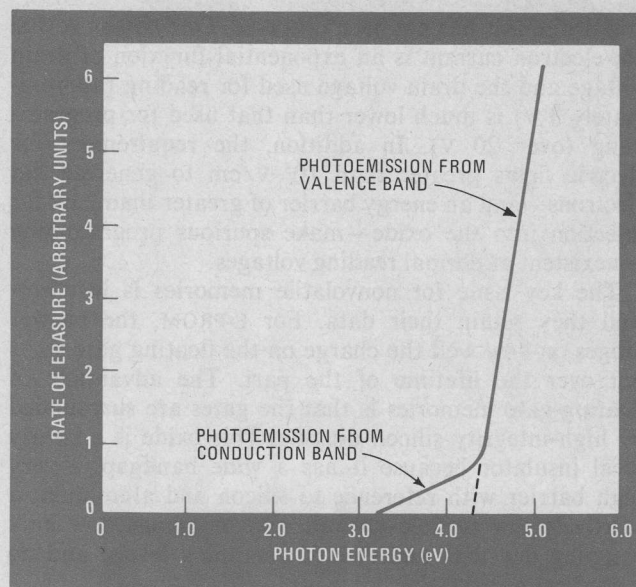


4. Injection. With a positive floating gate, electrons leave the drain and are accelerated along the channel. They gain enough energy to hop the thin oxide and cling to the attractive floating gate (as implied by the fields shown). This avalanche injection is self-limiting.

from the valence band of the polysilicon floating gate than from the conduction band.

The quantum yield and the erasure rate per incident photon follows the square-law dependence upon photon energy shown in Fig. 5. Two distinct threshold energies are apparent. The first, at 3.2 eV, is associated with the photo-excitation of electrons from the conduction band. Its slope is much shallower because of the lower density of electrons. The 4.3-eV threshold corresponds to the onset of photoemission from the valence band. The much steeper slope is indicative of the much higher density of valence-band electrons.

Unlike earlier p-type floating-gate E-PROMs, the fact that there is a significant density of electrons on the floating gate introduces the potential reliability problem of spurious erasure. Even with atmospheric filtering, sunlight photons having energies as high as 4.1 eV can reach the earth's surface. Also, fluorescent and incandescent lamps emit a minimum wavelength of about 3,000 Å, which corresponds to 4.1 eV. Erasure rates in sunlight or normal room light are much slower—1,000 times or more—than erasure under intense UV exposure because of the lower intensity of illumination and low density of



5. Erase rate. Two distinct threshold energies are apparent when erasing an E-PROM cell. The first, at 3.2 eV, corresponds to photoemission from the conduction band of the floating gate. The second point, at 4.3 eV, is associated with the valence band.

electrons in the conduction band of the floating gate. However, care must still be taken so that programmed E-PROMs are not subjected to prolonged exposure from either sunlight or ordinary room lighting.

Ultraviolet rays are so strongly absorbed that they do not get past the top select gate and the photons can only make their way to the floating gate from the side. In modern self-aligned structures, the edge of the floating gate is directly exposed to the radiation. But in earlier n-channel structures in which the floating gate completely overlaps the select gate, erasure is accomplished as the photons travel through the field oxide, under the select gate, and to the floating gate where they are absorbed. This wave-guiding effect is efficient because the reflectivity of silicon to 2,500-Å light is roughly 65%.

An interesting exception exists for very energetic X rays and gamma rays, which can readily penetrate the select gate and then be absorbed throughout the gate oxides, the floating gate, or the substrate. However, a broadband flux on the order of 10^{13} photons per cm from an X-ray tube operated at 100 kiloelectronvolts with a tungsten target is necessary to fully erase an E-PROM; hence X rays are not regarded as a special reliability concern for E-PROMs.

Finally, single ionizing events due to the passage of a cosmic ray or an alpha particle cannot cause spurious erasure because not enough charge is created in the gate oxides around the floating gate to neutralize the charge stored there.

Reading and storage

During reading, the E-PROM cell operates like an ordinary transistor except that the normal gate capacitance is replaced by the series capacitance of the floating-gate structure. Though it might be suspected that some low-level programming could occur during a read operation, as read voltages are applied to both the gate

and drain, this has not been observed. One reason is that hot-electron current is an exponential function of drain voltage and the drain voltage used for reading (approximately 3 V) is much lower than that used for programming (over 20 V). In addition, the requirement for electric fields greater than 10^5 V/cm to generate hot electrons—and an energy barrier of greater than 3 eV for injection into the oxide—make spurious programming nonexistent at normal reading voltages.

The key issue for nonvolatile memories is just how well they retain their data. For E-PROM, the answer hinges on how well the charge on the floating gate stays put over the lifetime of the part. The advantage of floating-gate memories is that the gates are surrounded by high-integrity silicon dioxide. This oxide is a nearly ideal insulator because it has a wide bandgap, a very high barrier with reference to silicon and aluminum, a relatively low surface-state density on silicon, low bulk trapping densities and electron trapping levels, and no structural polarization.

For ordinary electric fields from the floating gate to

the select gate and substrate encountered during reading and storage, electron emission is negligible. Therefore any charge leakage that occurs must be due to oxide defects. Since this leakage is due to a thermally activated hopping mechanism having an activation energy of 0.6 eV¹, its rate can be accelerated under high-temperature biasing and/or baking conditions. For example, programming all bits and then subjecting the device to a high temperature (150°C) rapidly determines the ability of each floating gate to maintain charge. Additionally, a high-temperature biasing of unprogrammed bits screens out oxide defects that cause current leakage to the floating gate.

By taking advantage of these high-temperature and/or voltage accelerating factors, production screens have been developed that successfully weed out those parts with defective oxides. These screens are the subject of part 2 of this article. □

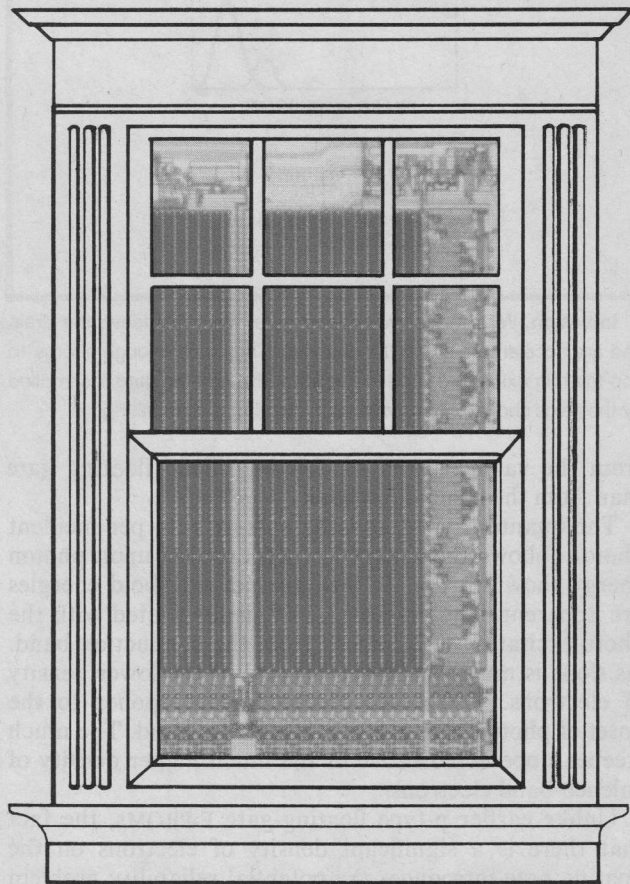
References

1. R. E. Shiner, J. M. Caywood, and B. L. Euzent, "Data Retention in E-PROMs," 18th Annual Proceedings, Reliability Physics Symposium (1980).

E-PROM reliability: part 2

Tests and screens weed out failures, project rates of reliability

by Stuart Rosenberg, *Intel Corp., Phoenix, Ariz.*



□ To determine the failure rate of a group of erasable programmable read-only memories—or any integrated circuit, for that matter—various short cuts must be employed or the testing time would become intolerable. Some E-PROM failure modes only occur after the device has been working properly for years (proper operation of an E-PROM is described in part 1 of this article). Obvi-

ously, production testing cannot afford this kind of time, so the aging process is accelerated with high temperatures and extreme supply voltages. This form of testing, called screening, will both pinpoint bad chips and help determine the overall failure rate that can be expected for a batch of devices.

The purpose of reliability testing is to quantify the

expected failure rate of a device at various points in its life. Fundamental principles of reliability engineering predict that the failure rate for a group of devices will follow the bathtub-shaped curve depicted in Fig. 1. This curve is divided into three regions: infant mortality, random failures, and wear-out failures. All classes of failure mechanism can be assigned to one or more of these regions.

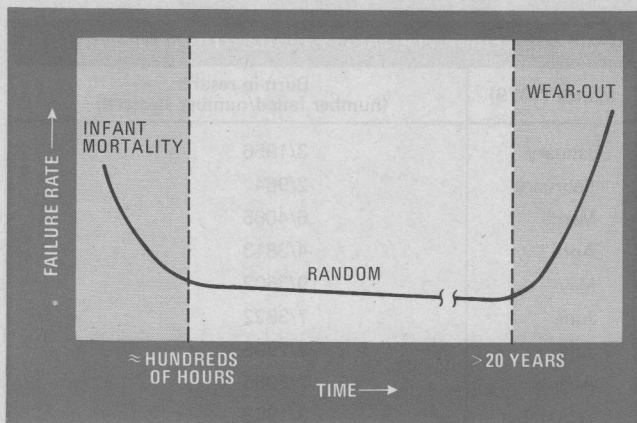
Infant mortality, as the name implies, represents the earliest failures in a device's life and is usually associated with one or more manufacturing defects. After a time, usually in the high tens to low hundreds of hours, the failure rate approaches a constant low value where it remains for a period of hundreds of thousands to millions of hours (for integrated circuits), depending upon temperature, applied voltages, and other operational and environmental factors.

This random-failure region of the curve represents the useful life of the device. During this period there is a slight decline in the failure rate as potential random failures are weeded from the general population. Wear-out failures are characterized by a rapidly rising failure rate with time as the devices wear out both physically and electrically.

Characterizing a given device type in the shortest possible time requires tests that simulate accelerated aging. To choose the right acceleration factor, it is first necessary to understand what makes E-PROMs fail in each region of Fig. 1.

Table 1 lists the common failure mechanisms for n-channel MOS E-PROMs. The table also shows the region of the bathtub curve most affected by each mechanism, as well as the thermal activation energy (E_a) of the mechanism. The thermal activation energy indicates the effect that increased temperature has on the frequency of the failure; the higher the activation energy, the greater the effect. For example, a temperature rise from 70°C to 125°C increases the failure rate resulting from oxide-hopping conduction ($E_a = 0.6$ electronvolt) by a factor of 16.5. This same temperature rise causes oxide rupture failures ($E_a = 0.3$ eV) to increase by only a factor of 6.

A detailed explanation of each mechanism is beyond the scope of this article, but charge loss, oxide-hopping



1. Failure tub. The reliability of an integrated circuit follows a bathtub-shaped curve. Infant failures happen right after the chip is made. Wear-out failures do not start until after 20 years, a time span representing the useful life of the device.

conduction, and hot-electron injection are responsible for the majority of failures. A buildup of electrons trapped in the oxide layer mentioned in part 1 commonly causes devices to wear out.

Charge loss from the floating gate is due to thermal emission. It affects most E-PROM cells to some degree over their 20-year life, and typically amounts to less than 10% of the charge stored on a cell.

Failure modes

Oxide-hopping conduction between the floating gate and the silicon substrate is typically less than 10^{-19} ampere, or about one electron per second. The electrons flow onto or off the floating gate depending on bias conditions, so they may cause bit errors on either stored 0s or 1s. Oxide-hopping conduction is related to manufacturing defects and does not occur in most devices. Thus, this failure mechanism affects the random and infant portions of the bathtub curve most strongly.

Undesirable hot-electron injection occurs when an MOS transistor is in saturation. Accelerated electrons in the pinch-off region may be scattered or attracted because the gate field is perpendicular to the silicon-oxide interface.

As covered in part 1 of this article, cell programming

TABLE 1: E-PROM FAILURE MECHANISMS

Mode	Lifetime region affected (in Fig. 1)	Thermal activation energy (eV)	Primary detection method
Slow trapping	wear-out	1.0	high-temperature bias
Surface charge	wear-out	0.5–1.0	high-temperature bias
Contamination	infant/wear-out	1.0–1.4	high-temperature bias
Polarization	wear-out	1.0	high-temperature bias
Electromigration	wear-out	1.0	high-temperature operating life
Microcracks	random	—	temperature cycling
Contacts	wear-out/infant	—	high-temperature operating life
Silicon defects	infant/random	0.3	high-temperature operating life
Oxide breakdown/leakage	infant/random	0.3	high-temperature operating life
Hot-electron injection	wear-out	—	low-temperature operating life
Fabrication defects	infant	—	high-temperature burn-in
Charge loss	infant/random/wear-out	1.4	high-temperature storage
Oxide-hopping conduction	infant/random	0.6	high-temperature storage/burn-in

TABLE 2: INFANT MORTALITY EVALUATION FOR THE 2716

Month (1979)	Burn-in results (number failed/number stressed)
January	3/1956
February	2/964
March	6/4085
April	4/3813
May	9/3602
June	7/3822
July	9/2995
August	7/3055
September	3/1893
October	8/5025
November	2/1726
December	13/3321
Total	73/36,257 = 0.2% failure rate
Failure analysis: 33 charge-lose failures 34 charge-gain failures 3 decoder failures due to oxide breakdown 2 ac failures from slow trapping or contamination 1 dc parametric failure from contamination	

of an E-PROM is accomplished through hot-electron injection. During a programming cycle, as electrons travel to the floating gate, a small percentage of these carriers may become trapped in the thin gate oxide. Since ultraviolet radiation is relatively ineffective at removing these trapped electrons during erasure, repeated programming and erase cycles may result in a build-up of charge. These trapped electrons may, in turn, cause significant degradation in the operating margins of the device and, ultimately, device wear-out.

Measuring mortality

To measure the infant-mortality failure rate, a 48-hour dynamic burn-in at 125°C is used. Depending on the failure mechanism, the test correlates with 240 to 2,000 hours of operation at 55°C. To perform such a test, E-PROM samples are gathered on a weekly basis, programmed with a checkerboard pattern, and burned in. Results of infant mortality tests done on the Intel 2716 E-PROM for 1979 are shown in Table 2. The data reveals an average infant-mortality failure rate of 0.2% for the year, with failures heavily dominated by hopping conduction (67 of a total of 73 failures).

To best characterize the random portion of the bathtub curve, burned-in E-PROMs that no longer exhibit infant mortality are used. Long-term failures are measured with 1,000-hour 125°C life tests; results for the 2716 are shown in Table 3, along with the failure analysis. There were 23 failures in 5.2×10^6 device hours for a failure rate of 0.4% per 1,000 hours at 125°C.

With more than one failure mode, and with different thermal activation energies, the effect of operation at 125°C varies for different devices. It should also be pointed out that since failures occur at the die level, any extrapolations must be made from junction, and not ambient, temperatures. Thus a 125°C to 55°C ambient

TABLE 3: 2716 DYNAMIC LIFE TEST DATA AND RANDOM-FAILURE RATE EVALUATION

Month started	Results at 125°C			
	168 hours	500 hours	1,000 hours	
Before 5/79	5/3983	7/3874	3/3969	
5/79	0/128	1/128	1/127	
6/79	0/128	2/128	0/126	
6/79	0/286	0/286	0/286	
6/79	0/29	0/29	0/29	
6/79	0/299	0/299	0/299	
7/79	0/192	0/192	0/192	
8/79	0/64	0/64	0/64	
8/79	0/99	0/99	0/99	
8/79	0/102	0/102	0/102	
8/79	0/102	0/102	0/102	
8/79	0/100	0/100	0/100	
8/79	0/100	0/100	0/100	
8/79	0/128	0/128	0/128	
9/79	0/64	1/64	0/63	
10/79	0/75	0/75	0/75	
10/79	0/31	0/31	0/31	
Cumulative	5/5894 (A)	12/5065 (B)	6/5055 (C)	
FAILURE ANALYSIS				
(A) 5 single-bit hopping conduction (0.6 eV)				
(B) 11 single-bit hopping conduction (0.6 eV) 1 double-bit charge retention (0.6 eV)				
(C) 5 single-bit hopping conduction (0.6 eV) 1 single-bit fabrication defect (0.3 eV)				
FAILURE RATE CALCULATION				
Failure-mode activation energy (eV)	Actual device-hours at 125°C	Equivalent device-hours at 55°C	Number of failures	Failure rate at 60% confidence level (% per 1,000 hours)
0.3	5.2 x 10 ⁶	2.7 x 10 ⁷	1	0.007
0.6	5.2 x 10 ⁶	2.2 x 10 ⁸	22	0.011
Total failure rate: 0.018% per 1,000 hours				

derating corresponds to a derating of junction temperatures from 130°C to 65°C for a 2716.

Although there are many ways to compute a 55°C failure rate, only one is realistic: by simply calculating the rate due to each mechanism and then adding them up. In another method, one that yields overly optimistic results, a single thermal activation energy is chosen to represent all failures in the population. A common choice is 0.7 eV since this value is used in military testing procedures. Based upon the number of device-hours in Table 3 (5.2×10^6) and this activation energy, the acceleration factor from 130° to 65°C corresponds to 48:1. This acceleration factor was read off an Arrhenius plot (Fig. 2).

For this scheme the equivalent number of device hours at 55°C works out to 2.5×10^8 . At a 60% confidence level with the tabulated number of failures (23), the 55°C failure rate becomes 0.01% per 1,000 hours. The pitfall of this method is that the acceleration factor is determined independently of the failure types and their corresponding activation energies. Unless all the failures are

0.7 eV, which is not generally the case, the extrapolation has little meaning.

Accurate results are given at the bottom of Table 3, where a 55°C failure rate of 0.018% for 1,000 hours is calculated. The first method's failure rate, 0.01%, is two times too optimistic.

Testing for wear-out

There are two failure mechanisms by which E-PROMs wear out that significantly affect reliability. These are charge loss and stress from program-erase cycling.

Since charge loss has a high thermal activation energy (1.4 eV), a test involving storage at a very high temperature (250°C) can determine where on the bathtub curve devices begin to wear out. So units with a high percentage of programmed bits (typically more than 90%) are placed in storage at 250°C for up to 1,000 hours. Intermediate data is taken to determine the shape of the failure-rate curve.

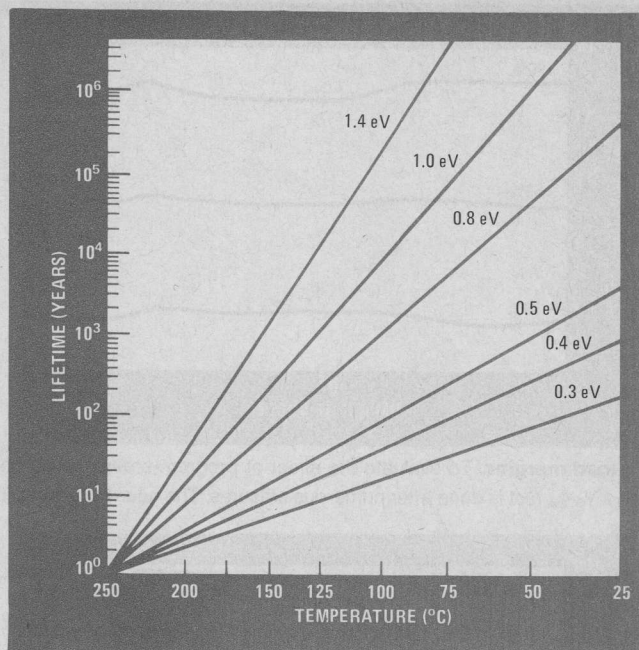
Figure 3 illustrates the cumulative percentage of failures versus time for the 2716 after a storage test at 250°C. This graph is based on data from more than 1,000 units, taken over a period of more than six months. This curve shows no sign of an increase in failure rate, indicating that the failures are defect-related and symptomatic of the infant and random failure regions. No indication of wear-out resulting from intrinsic charge loss is seen at 1,000 hours at 250°C. This corresponds to 1,950 years at 125°C or 1.3 million years at 70°C.

To evaluate the effect of program-erase cycling on device wear-out, supply voltage maximum and minimum measurements (V_{CCmax} and V_{CCmin} , respectively) are made after subjecting an E-PROM to multiple programming and erasing cycles. V_{CCmax} is the greatest supply voltage that will support proper operation, and V_{CCmin} is the lowest value. Each cycle comprises programming of the E-PROM with more than 90% 0s; measuring V_{CCmax} ; erasure via UV light with an intensity of 10 watt-seconds/square centimeter; programming to a pattern with more than 90% 1s; measuring V_{CCmin} ; and erasure.

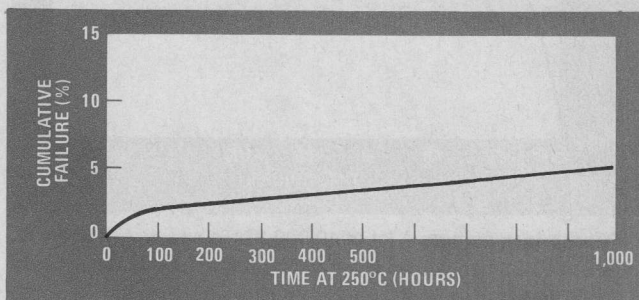
The results of 225 program-erase cycles on 48 2716s showed only one failure, which occurred after 56 cycles. The cause of the failure was determined to be partial oxide breakdown due to high field stressing during programming. Figures 4a and 4b show the 10%, 50%, and 90% points for program and erase margins (the cell's threshold voltage after programming and erasing) versus the number of program and erase cycles for the remainder of the sample. No appreciable shift can be seen in either the program or the erase margin and, for the 2716, degradation of operating margins due to hot electron injection did not occur.

Achieving acceptable reliability

To produce reliable E-PROMs, it is necessary to know the types of failures that may afflict the device and then devise tests to weed out those parts with latent defects. These tests are called failure-mode screens because they are effective only on those E-PROMs with latent defects due to a particular failure mechanism or mode. Figure 1 shows that infant mortality and wear-out failures each have an order of magnitude greater effect on reliability



2. After Arrhenius. The Arrhenius plot shows the extent that an increase in temperature will accelerate a failure mechanism. Using the thermal activation energy for the failure mode of interest—see Table 1—a device's life is deduced from a temperature differential.

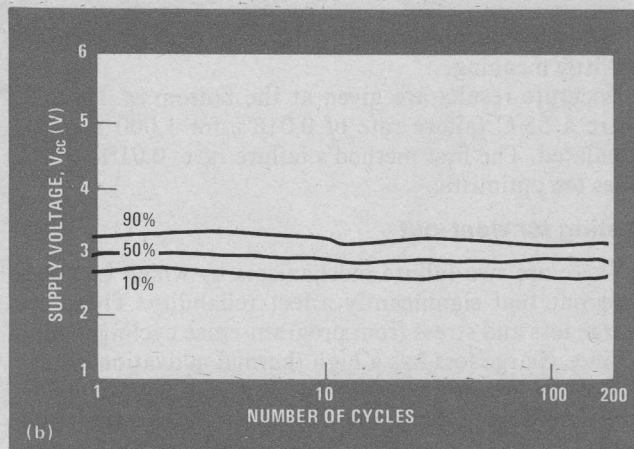
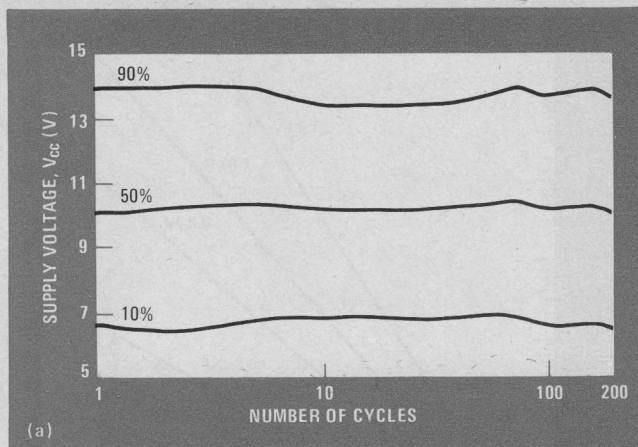


3. In sum. The cumulative percentage of failures is charted against time after storage tests done at 70°, 100°, and 250°C. The data was gathered from more than 1,000 units over a period of more than six months. Charge loss is not seen for an extended period of time.

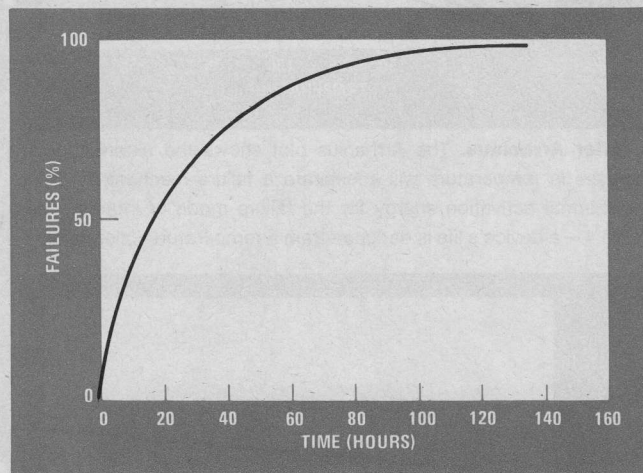
than do random failures. For this reason, failure-mode screens are designed to isolate mechanisms that strike in these two regions of the bathtub curve.

Screens are also region-specific, so two different screens are needed to reduce failures during both the infant mortality and wear-out periods. Although the infant-mortality screen does not lower the random-failure rate, it does reduce the cumulative failure rate, thereby improving overall reliability.

Infant-mortality screens reduce a population's early-life failure rate by prematurely aging each part just beyond the first change in slope of the bathtub curve. If the manufacturer can eliminate most parts with latent defects before they leave the factory, the balance of each shipment should be error-free until the parts begin to wear out. Wear-out screens attempt to extend the random-failure portion of the curve further to the right, extending useful device life. Infant-mortality screens have a negligible impact on the wear-out failure rate because of the relative time scales; premature aging of a



4. Good margins. To evaluate the effect of program-erase cycling on wear-out, a V_{CCmax} test is done after multiple programming cycles (a), and a V_{CCmin} test is done after numerous erasures. The near-horizontal lines indicate no degradation in program or erase margins.



5. Charge-loss screen. A lot of 10,000 2716s was screened for 72 hours at 150°C, then burned in for 48 hours to check for charge-loss failures. The optimum duration for such a screen corresponds to the curve's leveling out; this happens after 72 hours.

part—even by two months—means little after 20 years.

When faced with a raw population of E-PROMs whose reliability is to be improved by screening, the first step is thorough reliability life testing, followed by a detailed analysis of the failures. Once the dominant failure mechanisms have been identified and a failure rate computed, the number of screens and their types must be settled on.

Avoiding screen overkill

Screens cost money, though, because of the time and equipment needed to implement them. Overkill is a measure of the number of parts that fail a screen but that would never have failed in the field. A good screen has little overkill yet still lowers the failure rate. Ideally, screens should be part of a normal test program and should take less than a second to perform. A screen may be put into production to combat a new failure mode with little regard to overkill or complexity. Once the failure rate is under control, the screen can be modified to optimize its cost with respect to its effectiveness in improving reliability.

Two infant-mortality screens—one for defect-related charge loss and another for charge gain (both are mani-

festations of oxide-hopping conduction)—were developed for the 2716. The combined objective of those screens was to reduce the total early-life failure rate to less than 0.2% of the population. To measure the effectiveness of the screens, it was decided to follow each with a dynamic burn-in at 125°C for 48 hours—a typical infant-mortality period.

Charge-loss screen

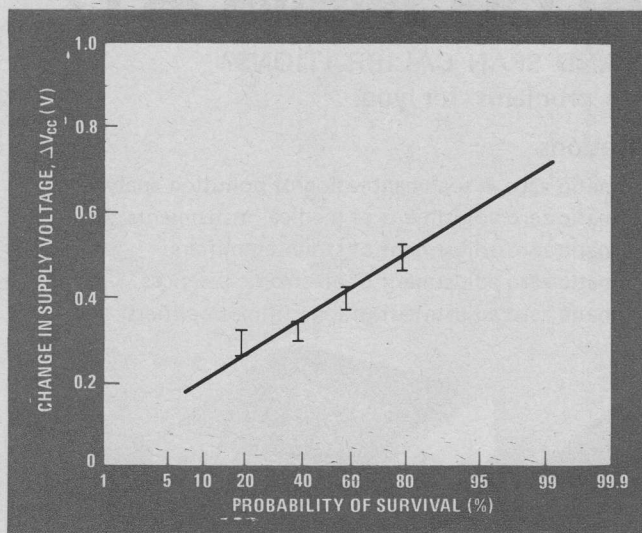
The charge-loss screen was devised only after an analysis of indicators of latent manufacturing defects, as these are most responsible for charge-loss failures. Though indicative of the number of electrons on the floating gate, ease of programming and erasure and programming margin were evaluated—but without success, in this case.

The next step was to examine the effect on charge loss of bias voltage and temperature. Earlier work with E-PROMs showed charge loss to be more than two orders of magnitude greater at 150°C than at 55°C, so a high-temperature bake was selected as the basic screen. To determine whether bias voltage affected the charge loss, programmed cells with and without bias were compared, but found to be similar. Therefore, it was decided to make the screen an unbiased high-temperature bake, for reasons of simplicity.

The final step was to determine how high in temperature and how long in duration the screen should be. Since it was intended for production runs of 2716s in tin-plated ceramic packages, temperatures were limited to 150°C to prevent severe lead oxidation. To determine the optimum bake time, the screen was applied to known charge-loss rejects with pass/fail tests every 20 hours. The results (Fig. 5) indicate that about 72 hours would be sufficient to weed out the majority of charge-loss rejects.

To confirm this time estimate, a 10,000-piece lot of 2716s was first screened for 72 hours at 150°C and then dynamically burned in for 48 hours. Approximately 0.1% of the parts that passed the charge-loss screen failed the burn-in. This percentage of charge-loss failures compared favorably with an overall infant-mortality failure rate of 0.27%.

Unlike charge loss, a well-known failure mechanism in



6. Wear-out. Wear-out screening should guarantee at least 99% of all parts will not fail during a specified lifetime. Both elevated temperature and maximum V_{CC} were used to screen for charge-loss wear-out. For a 1% failure rate, V_{CC} had to be increased 0.7 V.

older designs, charge gain seemed unique to advanced E-PROMs. Earlier work at Intel hinted at a relationship between supply voltage and charge gain, so a biased, dynamic burn-in for 48 hours at 125°C became the first screen attempted.

The temperature was later raised to 150°C when it was decided to run the screen on 2716s in their power-down mode. During power-down, all select gates are driven to V_{CC} and are stressed more effectively. By running the screen in power-down, the burn-in temperature could be safely raised to 150°C.

The supply voltage was the final parameter to be determined. Since the failure rate increases linearly with the stress voltage for charge gain, a V_{CC} level greater than 5 volts was desirable for the screens. But a conflict arose because the power-down mode at an elevated V_{CC} supply is also the programming mode. Fortunately, computer simulations of the 2716 revealed that 7.5 V could be used without going into the programming mode. Because of the increased voltage, temperature (150°C), and duty cycle, the screening time was reduced to six hours. The final result: a charge-gain screen done at 150°C for six hours at 7.5 V in the power-down mode.

With screens devised for both charge gain and charge loss, data was collected and measured against the 0.2% infant-mortality goal. Samples pulled weekly from an inventory awaiting shipment were subjected to a dynamic 48-hour burn-in at 125°C. Results for 1979 showed 72 failures out of 35,783 units stressed, or an infant mortality rate of 0.2%.

Wear-out can have a dramatic impact on reliability if it occurs in an operating piece of equipment. To minimize wear-out resulting from intrinsic charge loss, another type of screen was called for. As opposed to infant-mortality screening—where a small freak population must be eliminated—wear-out screening should insure that at least 99% of all parts will not fail during a specified lifetime.

It was known that temperature aggravates charge loss

TABLE 4: RESULTS OF DATA-RETENTION MONITOR

Month	Number of devices	Data-retention failures	% fail
1/78	293	14	4.7%
2/78	200	0	0%
3/78	300	4	1.3%
4/78	100	3	3.0%
5/78	300	1	0.3%
7/78	165	1	0.6%
8/78	260	3	1.1%
9/78	323	11	3.4%
10/78	176	4	2.2%
11/78	323	11	3.4%
12/78	535	9	1.6%

This test was run for 168 hours at 250°C

and that no external bias is necessary to induce charge loss in programmed cells. A good method for identifying charge-loss wear-out was found to be a 168-hour bake at 250°C. Using known acceleration factors from Arrhenius plots, this correlates with more than 220,000 years at 70°C. The acceptable percentage of failures for this test is 5%, which corresponds with 2×10^{-4} % failures due to charge loss in 10 years. If this rate is not exceeded, then wear-out has not yet begun.

Monitoring margins

Through the logging of cell programming margins (based on V_{CCmax}) at the beginning and end of the 168-hour bake at 250°C, an average degradation was determined. This was fed into the standard manufacturing test sequence—where cell programmability is checked—to insure that all parts have sufficient margin to weather a normal lifetime.

To determine a value for V_{CCmax} , experiments were conducted to discover the change in V_{CCmax} for 99% of the 2716s tested. Although the target for failures was 5%, a 1% ceiling was considered more appropriate to account for normal statistical variations in the small 50-piece sample.

The changes in V_{CCmax} for the 168-hour 250°C bake are plotted in Fig. 6 against the probability of seeing a given change in V_{CC} . From this plot, a ΔV_{CC} of 0.7 V corresponds to the 99% pass criterion. Thus, to insure that all 2716s have sufficient margin to withstand normal life, the standard manufacturing testing sequence includes a test with ΔV_{CC} set to +0.7 V.

By incorporating the screen in the normal product flow there is a minimum of disruption and no increase in testing time (the screen takes less than 1 millisecond to perform). The final test of this wear-out screen is its effectiveness. Results through the first 11 months of 1979 pinpointed 100 failures in 3,800 devices stressed. This works out to a failure rate of 2.5%, well below the target of 5% (see Table 4). □